Predicting Wind Energy Production in South Africa Using Machine Learning

A Research Report submitted in partial fulfilment of the Degree of Master of Commerce (Economics/Economic Science) in the School of Economics and Finance, University of the Witwatersrand

Student: Sivasha Reddy

Student Number: 2106536

Student Email: 2106536@students.wits.ac.za

Supervised by Professor Dambala Kutela

Research Paper Code: ECON7067A

Master's Programme Code: CCA01 (CFAEE60)

This Research paper counts for 50% of the Master's Programme

Word Count: 13 491 words (inclusive of references)

19/06/2025

Abstract

South Africa faces an urgent and escalating energy crisis driven by ageing coal infrastructure, frequent load shedding, and rising electricity demand. Wind energy presents a viable renewable energy alternative with significant potential to alleviate these challenges; however, its inherent variability complicates grid stability and energy planning. Accurate wind energy forecasting is essential for optimising power dispatch, minimising curtailment, and enhancing energy security. Despite advancements, traditional forecasting methods, such as physical models and statistical techniques, struggle to capture the complex and nonlinear nature of wind patterns, particularly in data-scarce environments like South Africa.

This study investigates the application of machine learning models to improve wind power forecasting in South Africa, where data constraints and fluctuating meteorological conditions pose unique challenges. The research examines the effectiveness of machine learning in predicting wind energy production and assesses the role of explainable artificial intelligence techniques, such as SHapley Additive exPlanations, in enhancing model transparency and interpretability.

Using historical meteorological data and turbine performance records from a South African independent power producer, the study evaluates multiple machine learning approaches to determine their predictive performance. A comparative analysis of different machine learning models highlights the most reliable techniques for wind energy prediction.

The findings demonstrate that XGBoost outperforms Random Forest, Decision Tree, and K-Nearest Neighbour. Furthermore, the machine learning methods show a significant improvement over traditional statistical techniques, offering improved predictive accuracy while providing insights into key meteorological and operational factors influencing wind power generation. The integration of explainable artificial intelligence further ensures interpretability, fostering trust and practical usability among stakeholders.

This research contributes to the renewable energy forecasting literature by adapting machine learning solutions to a data-scarce environment and emphasising the role of interpretability in real-world adoption. The results provide valuable insights for policymakers, energy planners, and grid operators, supporting South Africa's transition to a more sustainable and resilient energy future.

However, the study's findings may be limited in generalisability and accuracy due to the analysis focusing on a single wind turbine chosen based on data availability rather than representativeness. Consequently, the results may not extend to other wind farms or turbines operating under different geographic or climatic conditions.

1. Introduction

Background and motivation

South Africa stands at a critical juncture in its energy history, grappling with the need to reduce carbon emissions, satisfy rising electricity demands, and stabilise an ageing and often unreliable grid. For much of its industrial development, the country has relied heavily on coal-fired power stations, building a power sector that, although historically robust in capacity, increasingly shows signs of significant strain. As illustrated in Figure 1, recurrent and escalating load shedding, coupled with mounting maintenance costs and escalating environmental pressures, underscore that the status quo is no longer tenable.

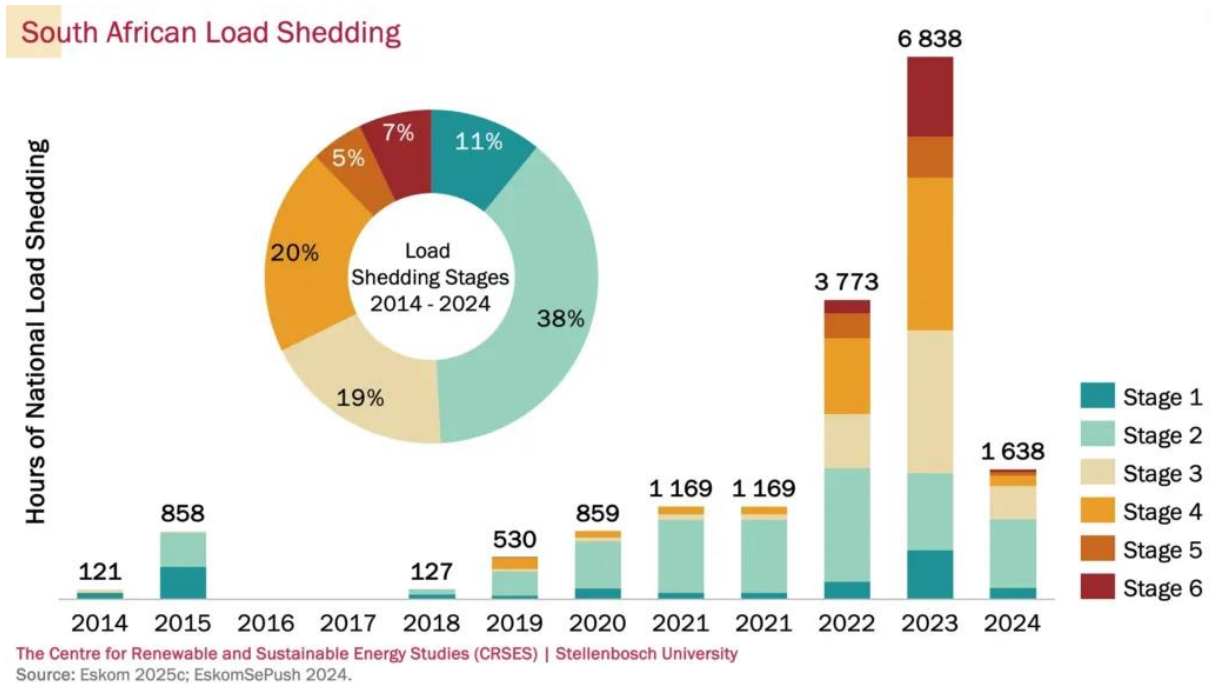

Figure 1 – South African Loadshedding Hours (CRSES, 2025)

Against this backdrop, the search for renewable alternatives has gained momentum. Wind power has emerged as one of the most promising prospects, bolstered by government initiatives designed to catalyse private investment and by the sheer abundance of wind resources found along South Africa's coastline and certain inland areas (see Figure 2).

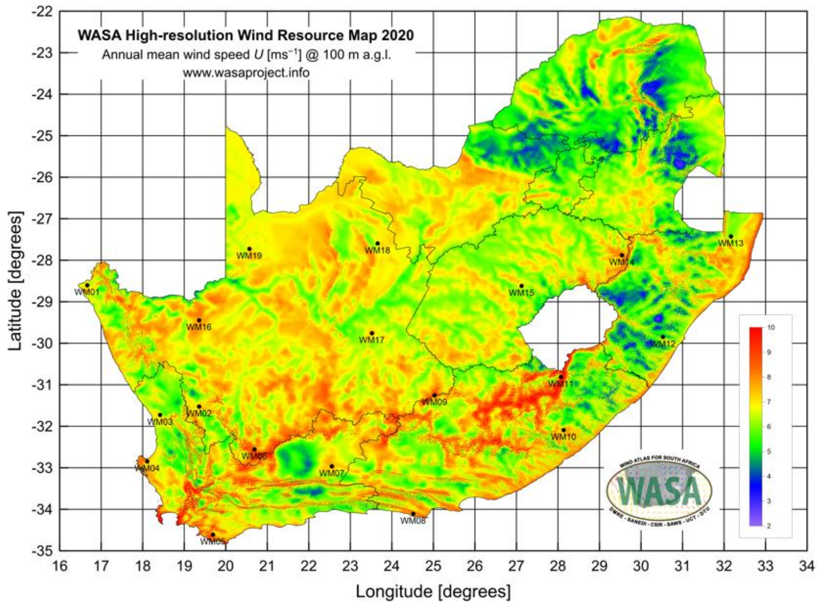

Figure 2 – South Africa Wind Resource Map (South African Wind Atlas, 2020)

As shown in Figure 3, South Africa remains heavily reliant on coal, which constitutes the majority of its energy production. The industrial sector is the country's largest energy consumer, underscoring the vital role of the energy sector in sustaining economic growth (Cowling, 2024). However, this dependence on a non-renewable and environmentally harmful resource highlights the urgent need to diversify the energy mix while enhancing the efficiency and reliability of energy production and distribution.

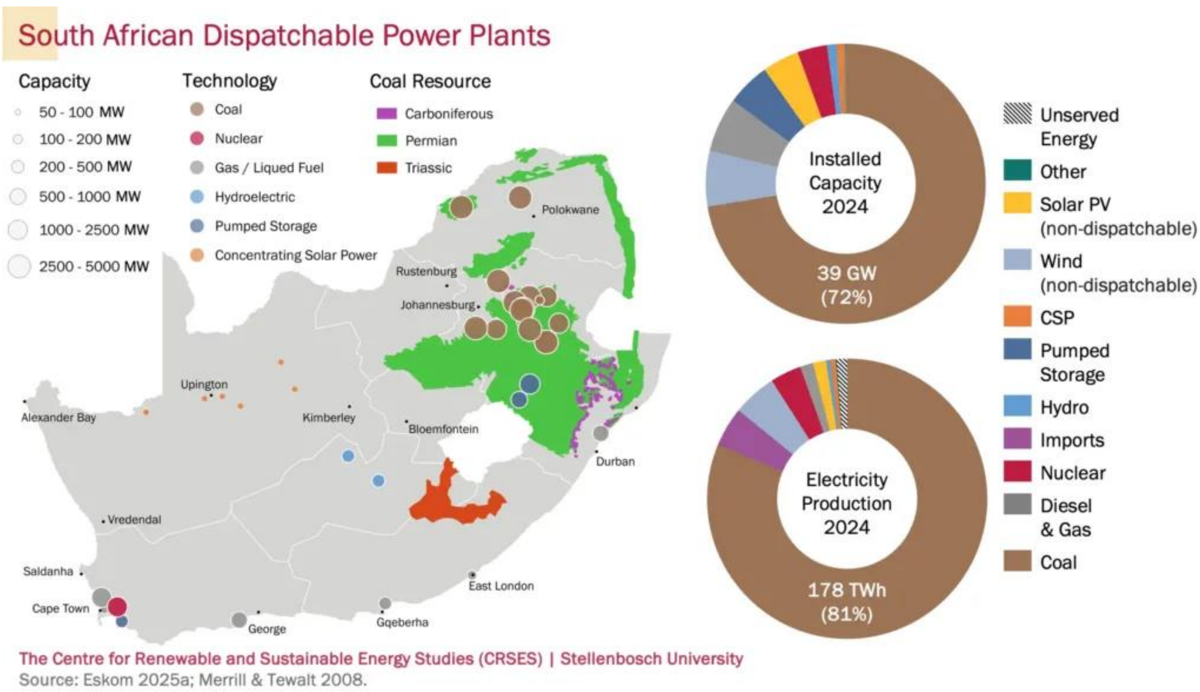

Figure 3 – South Africa's Energy Mix (CRSES, 2025)

The integration of renewable energy sources, particularly wind power, has become essential for reducing environmental impacts and strengthening energy security. Figure 3 indicates that wind energy is the second-largest source of electricity generation overall and the leading renewable energy contributor in the country. However, for wind energy to effectively displace a portion of coal-generated electricity and provide consistent, reliable power, it must overcome critical hurdles. These include accurately forecasting wind speeds and energy output, ensuring comprehensive integration with existing infrastructure, and clear communication of benefits and potential risks to all stakeholders.

Since the early 2010s, South Africa's Renewable Energy Independent Power Producer Procurement Programme (REIPPPP) has been instrumental in diversifying the country's energy mix by accelerating the adoption of renewable energy sources. This initiative has successfully expanded the nation's renewable energy capacity, attracting significant investment, both locally and globally, from investors who are interested in clean energy technologies (Eberhard & Naude, 2016). By late 2023, wind power projects approved under the REIPPPP had contributed over 6,000 megawatts of installed or contracted capacity, with additional procurement rounds planntimeseries

ed to further strengthen the sector.

Despite this growth, Figure 3 illustrates that wind energy production still represents a relatively modest share of South Africa's overall energy mix, constrained by the inherent variability of wind patterns and structural challenges within the national grid. While the country's coastal corridors provide a relatively stable wind resource, periodic fluctuations in wind speed and direction introduce volatility in daily generation volumes.

Grid operators, already burdened by routine power shortages (see Figure 1), face additional challenges when attempting to integrate intermittent wind power without advanced forecasting tools. Although methods such as day-ahead and hour-ahead scheduling exist, inaccurate predictions can lead to curtailment, unplanned outages, or an overreliance on fossil-fuel-based backup systems, each of which undermines the environmental and economic gains that wind promises.

Efforts to address these challenges have traditionally followed two main approaches. The first relies on physical or numerical weather prediction (NWP) models, which integrate meteorological equations with geospatial data on terrain and wind turbine characteristics (Lange & Focken, 2006). While these models are theoretically robust, they require substantial computational resources and dense observational networks, both of which are limited in many regions of South Africa.

The second approach involves a time series analysis of historical wind speed and energy production data, employing techniques such as autoregressive integrated moving average (ARIMA) models (Elsaraiti et al., 2019). While effective when historical data is stable and extensive, these time series methods often struggle with sudden changes in wind behaviour or incomplete datasets—challenges that are particularly prevalent in emerging economies like South Africa.

Research problem

Despite the promise of wind energy, multiple challenges obstruct its effective deployment in South Africa. According to Wright and Grab (2017), wind speeds are inherently volatile due to complex meteorological and geographic factors. South Africa experiences varying wind patterns due to coastal and inland dynamics, seasonal shifts, and localised turbulence, making accurate forecasting difficult. Consequently, forecasting models often fail to capture their full complexity, leading to inaccuracies in production estimates. These inaccuracies pose risks to grid stability, as sudden shortfalls or surpluses can strain the power system. Additionally, deviations in production estimates can escalate operational costs, necessitating the use of expensive backup energy sources, greater reliance on storage solutions, and an increase in overall costs, especially under constraints such as limited data availability, inconsistent meteorological patterns, and financial hurdles.

High-precision forecasting is critical for reliability and cost-effectiveness, yet existing methods frequently underperform when applied to South Africa's environment. Several constraints exacerbate these forecasting difficulties. Limited data availability, particularly in regions with sparse meteorological monitoring infrastructure, restricts model training and validation. Inconsistent meteorological patterns, influenced by sudden weather shifts and localised anomalies, reduce the reliability of traditional statistical and physics-based forecasting approaches.

Machine learning has gained traction as a potential solution to improve wind energy forecasting, offering the ability to identify hidden patterns in data and enhance predictive accuracy. However, existing machine learning-based forecasting methods frequently underperform in South Africa's unique environmental conditions. Many models are developed using datasets from other regions with different wind regimes, limiting their effectiveness when applied locally. Additionally, data scarcity and quality issues present further challenges, as many machine learning approaches rely on large, high-resolution datasets to perform optimally. A major barrier to the adoption of machine learning forecasting models is their lack of interpretability (Olawade et al., 2024). Stakeholders, including policymakers, grid operators, and investors, are hesitant to rely on models that function as "black boxes," providing predictions without clear explanations of their decision-making processes (Jha, 2023). The inability to interpret model outputs makes it difficult to assess reliability, troubleshoot errors, and integrate machine learning-based forecasts into real-world energy planning and policy frameworks. While techniques such as explainable artificial intelligence and feature attribution methods offer potential solutions, their application in wind energy forecasting in South Africa could be explored further.

Research Question

Given these constraints, the primary research question guiding this study is: How effective are machine learning models in predicting wind energy production in South Africa, considering the country's unique geographic and climatic conditions?

To address this overarching question, the study also examines three key sub-questions:

a. Which machine learning models yield the most accurate forecasts under local conditions, and how do they compare with traditional statistical methods?
b. What are the critical operational and meteorological variables influencing wind energy production, as identified by explainable machine learning techniques, and how can these insights guide decisions on turbine placement, maintenance strategies, and policy decisions?
c. How can model transparency be enhanced to encourage adoption by grid operators and policymakers, ensuring that advanced forecasting tools are both effective and trusted?

Hypothesis/tentative claim

This study posits that machine learning models, when carefully optimised for data constraints and coupled with interpretability methods, can substantially improve wind forecasting in South Africa. Traditional statistical techniques, including ARIMA, often struggle with nonlinearities and the dynamic interplay of meteorological factors influencing turbine performance. In contrast, machine learning can integrate a variety of inputs—such as wind speed, temperature, and turbine specifications—while dynamically adapting to changing conditions. While the inherent complexity of machine learning models can deter practitioners if interpretability is lacking, techniques like (SHapley Additive exPlanations) SHAP can uncover the rationales behind model outputs, thereby mitigating concerns about "black box" approaches.

Contribution to the Literature

This study contributes to both academic research and practical applications. Academically, it advances the understanding of machine learning–based wind forecasting in a data-constrained setting, offering insights into how to adapt algorithms to real-world complexities, such as limited meteorological records and incomplete turbine logs. By comparing diverse machine learning models with traditional approaches, the study clarifies their strengths and limitations in predicting wind energy under South Africa's unique conditions. Furthermore, it underscores the importance of model explainability, demonstrating how interpretable techniques, such as SHAP, can enhance stakeholder trust and inform policy decisions.

From a practical perspective, the findings provide valuable guidance for grid operators, policymakers, and energy planners on integrating wind power more effectively. The study's emphasis on interpretability identifies the meteorological and operational factors that most strongly influence production, facilitating optimal turbine placement and improved maintenance scheduling. These actionable insights not only support the growth of renewables but also help mitigate risks associated with load shedding and underutilised capacity.

Structure of Thesis

The remainder of this thesis is organised as follows:

Section 2 provides a comprehensive review of the existing literature on wind energy forecasting, spanning both traditional statistical methods and emerging machine learning approaches. This section also situates the study within relevant theoretical frameworks, highlighting the role of interpretability in bridging knowledge gaps.

Section 3 outlines the study's methodology, detailing data sources, preprocessing steps, and model selection criteria. It further explains how each forecasting technique is applied, how hyperparameters are optimised, and which evaluation metrics are used to evaluate performance.

Section 4 presents the empirical findings, comparing the accuracy and interpretability of various machine learning models under South African conditions. It delves into feature importance results, identifying the operational and meteorological factors that most strongly affect wind energy production.

Section 5 analyses the results and discusses the findings of the thesis with a focus on possible policy implementation and large-scale grid management.

Section 6 concludes the thesis by summarising key outcomes, acknowledging limitations, and proposing avenues for future research, particularly with respect to policy implementation and large-scale grid management. It reaffirms the significance of explainable machine learning solutions for the sustainable and efficient deployment of wind energy.

2. Literature Review

Statistical methods for wind energy forecasting can be categorised into three main sections: traditional time series analysis, machine learning models and deep learning approaches. While this paper focuses on the analysis of machine learning approaches for wind energy production forecasting, this section discusses prior literature on all statistical methods.

Time Series

Several studies have identified SARIMA as a reliable forecasting tool (Liu et al., 2021; Tyass et al., 2022; Szostek et al., 2024), with some demonstrating its superiority over machine learning techniques. Szostek et al. (2024) highlighted the strengths of both ARIMA and SARIMA while revealing that Support Vector Machines (SVMs) performed comparatively weaker. The limited amount of data used in their analysis, coupled with the absence of multi-dimensional inputs, likely contributed to the underperformance of SVM. Similarly, Liu et al. (2021) found that SARIMA outperforms the deep learning methods, Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) in forecasting accuracy for offshore wind speeds. However, the relatively small dataset (one year) and limited training period (one month) likely constrained the deep learning models' ability to effectively capture and learn the underlying patterns within seasonal periods.

Machine Learning

In a critical review of energy economics literature, Ghoddusi et al. (2019) highlighted that Genetic Algorithms (GAs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs) are among the most widely used techniques in energy economics research. However, their focus is on the broader energy economics field, rather than specifically on wind energy forecasting.

Dhungana (2025) provides evidence that XGBoost outperforms other machine learning models, including LR, KNN, and RF, as well as the following deep learning models: GNN, RNN, GRU, and LSTM, when considering computational resource efficiency.

In contrast, Maghraoui et al. (2022) conducted an empirical analysis which demonstrates the superiority of Random Forest over other machine learning models, such as ANN, Decision Trees (DT), and SVM. However, their study raises concerns about the lack of discussion on feature importance and model interpretability, which may affect the practical applicability of these models in real-world applications.

Similarly, Ledmaoui et al. (2023) found that ANNs outperformed other machine learning models, including Support Vector Regression (SVR), DT, Random Forest (RF), Generalised Additive Models (GAM), and XGBoost, in forecasting solar energy production, particularly when using GridSearchCV for hyperparameter optimisation. However, their study did not address model interpretability or the integration of multidimensional inputs, which are critical for applications like wind energy forecasting.

Khan et al. (2020) presented a machine learning-based hybrid approach that combined XGBoost, CatBoost, and Random Forest. Their proposed models outperformed the individual models, highlighting the potential of hybrid machine learning techniques to improve forecasting accuracy. However, while their models incorporated time-based features, they excluded other important factors, such as temperature, which could further enhance the accuracy of the forecasts.

Deep Learning

Numerous studies have highlighted deep learning models' superior performance to both machine learning and time series techniques (Li et al., 2023; Ledmaoui et al., 2023; Bousla et al., 2024).

Bousla et al. (2024) results demonstrated the Recurrent Neural Networks (RNNs) ability in wind power forecasting, highlighting their advantages over DNN and SVM models. However, the study was based on a relatively short dataset spanning only two years, which may not capture long-term seasonal variations or extreme weather events that can significantly impact wind energy production.

An empirical study compared time series models (ARIMA, GM), traditional machine learning models (LR, RF, SVR), and deep learning models (ANN, LSTM, CNN). Li et al.'s (2023) findings highlighted the superior performance of Convolutional Neural Network (CNN) in predicting wind speed. However, the study does not explore the detailed optimisation of model hyperparameters. Thus, the results may not fully reflect the potential of the machine and deep learning models, which are highly sensitive to hyperparameter tuning.

Zhou et al. (2021a) introduced a hybrid deep learning approach that outperformed MLP, LSTM, and ALSTM for photovoltaic energy forecasting. The study effectively utilised Convolutional Neural Networks (CNN) for feature extraction and an attention-enhanced Long Short-Term Memory (LSTM) model to prioritise critical features. However, it did not sufficiently address computational efficiency and scalability, and the focus was primarily on short-term predictions.

Wind Energy Forecasting in South Africa

Existing literature on machine learning for wind energy production in South Africa includes Thanyani's (2020) comparative analysis of machine learning models (SVM, Stochastic Gradient Boosting, and Neural Networks) for short-term electricity demand forecasting. Similarly, Daniel et al. (2020) explore short-term wind power forecasting in South Africa, focusing on the performance of ANN, SGB, and GAM.

Mhlanga (2023) provides an optimistic outlook on the potential of artificial intelligence and machine learning in addressing the challenges faced by the energy sector in developing countries. While the author highlights key issues such as rising consumption, inefficiencies, fluctuating demand, and the prevalence of unauthorised grid connections, the paper focuses primarily on how artificial intelligence and machine learning can be leveraged for predictive maintenance, energy consumption optimisation, and improved grid management.

Thus, while renewable energy forecasting in South Africa has received considerable attention in recent research, significant gaps remain in the forecasting of wind energy.

Limited Scope of Existing Studies

Prior literature explored short-term electricity demand forecasting using machine learning techniques (Thanyani, 2020; Daniel et al., 2020). Although short-term load forecasting is critical for operational efficiency, it falls short of addressing long-term energy planning needs.

Furthermore, while Ramarope et al. (2023) examined machine learning models in the renewable energy sector, the focus was specifically on hydroelectric generation forecasting. Hence, the findings may not be generalisable to wind energy.

Given South Africa's ongoing energy crisis, the absence of long-term wind energy forecasts and the lack of generalisability diminishes the relevance of these studies for sustainable energy policy development and grid stability.

Insufficient Model Transparency and Interpretability

A key issue in existing studies is the lack of model transparency and interpretability. For example, Makubyane and Maposa (2024) provided performance metrics such as AIC, BIC, RMSE, and MAPE but did not discuss model interpretability, which is crucial for adoption by grid operators and policymakers. Krishnamurthy et al. (2024) explored various AI models, including hybrid approaches, but did not address the "black box" nature of complex models like LSTM and CNN. Similarly, Thanyani (2020) discussed the trade-off between model accuracy and interpretability, noting that while techniques like SGB and ANN offer strong predictive performance, they reduce interpretability.

Model transparency is essential for ensuring adoption by grid operators and policymakers, yet it remains an overlooked aspect in many forecasting studies. The lack of transparency present could pose significant challenges for stakeholders in the energy sector who require clear insights into how predictions are made, especially in a developing context like South Africa.

Data Quality and Availability

Mhlanga (2023) underscores a critical issue prevalent in emerging markets: poor data availability. This is largely attributed to unauthorised grid connections and the lack of advanced metering infrastructure, which hinders accurate data collection and analysis. Such limitations pose significant challenges for researchers and policymakers aiming to derive actionable insights from energy data in South Africa.

The study conducted by Thanyani (2020) relies on a relatively limited dataset, consisting solely of historical hourly electricity demand data without incorporating additional variables. This narrow scope restricts the depth of analysis and limits the study's ability to account for broader influencing factors.

Furthermore, Makubyane and Maposa (2024) highlight persistent concerns regarding data availability and quality, a recurring challenge in studies of this nature. They argue that extending the historical dataset to span multiple decades would provide a more comprehensive foundation for analysis, thereby improving the robustness and generalisability of the findings. Addressing these data-related challenges is crucial for advancing research in the energy sector, particularly in contexts where data reliability and accessibility remain significant obstacles.

Contributions of this study

To address these gaps, this paper aims to:

- Focus on long-term forecasting to account for climate variability and assess the sustainability of wind energy integration into the national grid.
- Utilise high-resolution, localised wind energy datasets to enhance forecasting accuracy and reliability.
- Ensure model transparency and interpretability, guaranteeing that the forecasting models are both accurate and actionable for policymakers and energy grid operators.

By incorporating these advancements, the study aims to enhance South Africa's energy planning strategies and optimise the integration of wind energy into the national grid. A more comprehensive and transparent forecasting framework will support the development of sustainable energy policies and contribute to the country's long-term energy security. Unlike prior research, which has overlooked advanced interpretability techniques, this study employs model-agnostic methods to ensure both local relevance and robust explainability, thereby bridging critical gaps in energy forecasting and planning.

3. Methodology

Sample Selection

The dataset originates from an independent power producer (IPP) in South Africa and comprises 10-minute intervals of wind energy production data collected from a wind farm in South Africa. Spanning 10 years from the year 2015 to 2024, the dataset contains 490,140 data points and includes a range of features such as energy production, turbine mechanical performance, environmental conditions, and blade dynamics.

All turbines in the wind farm were mechanically identical. To ensure a more complete and reliable analysis, a single wind turbine was selected for this study, as it had the most comprehensive dataset among all turbines in the wind farm. However, it is important to acknowledge that this selection was based on data availability rather than the representativeness of South Africa's broader wind energy sector. Consequently, any findings may not be directly generalisable to wind farms with different turbine models, geographic conditions, or operational characteristics.

The data was collected through sensors, data loggers, and condition monitoring systems installed on the wind farm. Further detailed information about the wind farm was not disclosed due to confidentiality.

Feature Selection

The factors influencing wind power production are multifaceted and include several key variables. According to Dhungana (2025), these variables can be categorised as follows: temporal data, derived from timestamps, which provides insight into time-based patterns of energy production; meteorological data, including weather measurements such as wind speed and environmental temperature, which directly impact the efficiency of wind turbines; and power curtailment data, which comes from turbine signals such as blade pitch angle, rotor speed, and relative wind direction, all of which play a crucial role in optimising turbine performance and energy output.

The selection of independent variables was guided by their relevance to the predictive accuracy of the model. However, data availability posed a limitation, as missing variables such as air pressure and humidity could potentially influence model performance. The omission of these factors might introduce bias into the results, impacting the model's predictive capability.

Given the available data and feature engineering outlined in the following section, the following dependent and independent variables were included in the model. Table 1 summarises the variables included in the model, coupled with a brief description of the variables and their possible effect on energy production.

Table 1 Summary of variables

Data Types	Variable name	Variable description
Temporal	Datetime	The timestamp for each recorded measurement, including the 12-hour and 24-hour lags detects date and time for each 10-minute interval
	Hour of the Day	Categorical variable indicating the hour of the day
	Day of the Month	Categorical variable indicating the day of the month

	Day of the Week	Categorical variable indicating the day of the week
	Month of the Year	Categorical variable indicating the month
	Season	Categorical variable indicating the four seasons
Production	Energy Production (kWh)	The total energy generated by the wind turbine in every 10-minute interval, measured in kilowatt-hours.
	Average filtered wind speed (m/s)	Real-time filtered wind speed measurement with a 1-second interval, used for dynamic turbine control and safety responses.
Meteorological	Average environment temperature (°C)	The ambient temperature around the wind turbine helps to assess environmental effects on turbine performance and detect extreme conditions like freezing or overheating.
	Turbine age (months)	The age of the turbine in months
	Average rotor speed (rpm)	The average rotational speed of the rotor is measured in revolutions per minute (rpm). This is an indicator of power generation efficiency. Extreme speeds may reduce performance or cause mechanical stress.
Power Curtailment	Average angle Blade (°)	These represent the average pitch angles of each turbine blade. Adjusting blade pitch may help optimise the amount of energy captured in varying wind conditions and prevent damage during high winds.
	Average nacelle temp (°C)	The average internal temperature of the nacelle (the housing containing the generator, gearbox, and other components). High temperatures may indicate inefficiencies, overheating, or ventilation issues.
	Yaw duration (s)	The total time the yaw system has been actively adjusting the turbine's orientation to the left or right. A high value may indicate frequent wind direction changes, increased wear on yaw components, or inefficient tracking.
Interaction Features	Interaction of Wind Speed & Blade Angle	This feature captures the combined effect of wind speed and turbine blade pitch on energy production.
	Interaction of Nacelle & Environment Temperature	This feature captures the combined effect of nacelle temperature and environment temperature on energy production.

Rolling Mean & Standard Deviation Features	Wind Speed (m/s) Rolling Mean 24hr	The 24-hour rolling mean of wind speed smooths short-term fluctuations. This feature is useful for capturing daily wind patterns and seasonal variations.
	Wind Speed (m/s) Rolling Std 24hr	The 24-hour rolling standard deviation of wind speed quantifies wind variability over a full day.
Lagged Features	Wind speed (m/s) 12hr Lag Wind speed (m/s) 24hr Lag	The 12 & 24-hour lagged average filtered wind speed captures cyclical trends and delayed effects, and seasonal variations by reflecting how past wind conditions influence current energy production.
	Energy production (kWh) 12hr Lag Energy production (kWh) 24hr Lag	The 12 & 24-hour lagged energy production captures cyclical trends and delayed effects by reflecting how past energy production influences current energy production.

Data Preprocessing & Cleaning

Jupyter Notebook was utilised for developing, debugging, and visualising algorithms for this research. The Python libraries pandas, NumPy, statsmodels and scikit-learn were employed to facilitate data manipulation, model training, evaluation, and visualisation.

Pre-processing was employed on the dataset to address any potential errors, inconsistencies, or missing values in the raw data. This step ensured the reliability and accuracy of the training and testing datasets. If a model is trained on erroneous data, it will produce inconsistent predictions, leading to incorrect interpretations and flawed counterfactual explanations.

To address potential multicollinearity concerns, Variance Inflation Factor (VIF) checks were conducted for all predictor variables. After the removal of features with particularly high VIF values over 10, the results indicated that the remaining variables had VIF values below 6, suggesting mild multicollinearity. This ensured that predictor variables contributed uniquely to the model without redundant information, thereby improving the model's stability and interpretability.

An autocorrelation analysis was performed on all variables to assess the correlation between a time series and its lagged values, helping identify trends, seasonality, and long-term dependencies. The autocorrelation function (ACF) exhibited a gradual decay for all variables, indicating a slow decline in autocorrelation over time. The sustained positive correlations across the maximum lag range suggest persistent patterns driven by underlying trends or seasonal effects. This implies the presence of long-memory effects, where past values significantly influence future observations.

To evaluate the stationarity of the time series data, the Augmented Dickey-Fuller (ADF) test was conducted for each variable. The results yielded a p-value of 0, confirming that all variables are stationary with no unit roots detected. As a result, differencing was not required, indicating that the time series data was already stationary and did not need further transformation.

Outliers in the dataset were identified using the Interquartile Range (IQR) method, with a 1.5 IQR multiplier applied to detect extreme values. Identified outliers were replaced with NaN values to be handled appropriately during the imputation process.

The rolling median method was used to fill missing values (NaNs) in the dataset. This approach effectively smooths out noise and irregularities in time series data while preserving data integrity. The rolling median provides a robust imputation technique by leveraging the local context of surrounding data points. Its resistance to outliers ensures that extreme values do not disproportionately influence the imputed values. Furthermore, this method maintains underlying trends and patterns without excessive smoothing, making it well-suited for this dataset.

Feature Engineering

Time-based Features

Wind energy production is highly dependent on time-based variations that occur on daily, weekly, seasonal, and yearly scales. To effectively capture these patterns, several time-based features were generated from the timestamp data. The *hour of the day* feature helps identify daily patterns, as wind speeds and energy production often vary throughout the day due to changing atmospheric conditions. The *day of the week* feature accounts for potential weekly cycles that may arise from consistent environmental or operational factors. The *month of the year* feature and the *season* feature capture seasonal variations, such as differences in wind patterns between summer and winter. Additionally, the *week of the year* feature tracks broader seasonal trends, allowing for the detection of long-term variations in wind energy production. The inclusion of these time-based features is aimed at enhancing the model's accuracy and performance by identifying repeating cycles and seasonality of model features.

Rolling means and standard deviations

Rolling means and standard deviations smooth out short-term noise and capture overall trends. The inclusion of wind speed rolling mean and standard deviation is useful for detecting persistent increases or decreases in wind speed and identifying wind variability and turbulence.

Lagged Features

Wind speed (current and lagged values) is the most influential factor, as power output is proportional to the cube of wind speed, according to Betz's Law (Burton et al., 2011). Since wind patterns exhibit fluctuations, incorporating lagged wind speed values allows models to account for inertia and autocorrelation in wind changes, improving predictive accuracy.

Similarly, lagged energy production values are essential, as wind turbines have rotational inertia, meaning power output at a given moment is influenced by previous energy production. This feature enhances forecasting by learning transition patterns in energy generation over time.

The 12-hour and 24-hour lags detect daily cycles and seasonal changes by explaining how past values of wind speed and energy production impact energy production now.

Interaction Features

The interaction term of wind speed and blade angle is a feature that captures the combined effect of wind speed and turbine blade pitch on energy production. The blade angle can be dynamically adjusted based on wind speed to optimise energy capture and ensure safe operation (Chen et al., 2023). Including this interaction term allows for more accurate forecasting of energy production and optimising of pitch control algorithms, which may assist in identifying the optimal blade angle for varying wind conditions (Zhou et al., 2025b).

Feature normalisation

As part of the dataset's pre-processing, numerical features were normalised, and categorical features were transformed into numerical values. Normalisation ensured that all features were on a uniform scale, thereby preventing any individual feature from exerting undue influence on the machine learning process. This approach not only facilitated more rapid convergence of the model during training but also assisted in enhancing model performance (Richman, 2025).

Categorical features were encoded into numerical values to meet the requirements of machine learning algorithms. This transformation was essential for enabling algorithms to process and extract meaningful insights from categorical data. By converting categorical variables into numerical form, the risk of misinterpreting ordinal relationships or introducing bias among categories was minimised, thereby improving fairness and reducing errors. This step was crucial for ensuring the reliability of the trained machine learning models.

Feature normalisation was performed using the *MinMaxScaler* class from the *scikit-learn* library, scaling values to a range between 0 and 1. This prevented features with different original scales from exerting undue influence on the model. The impact of normalisation on the dataset is illustrated in Figure 4.

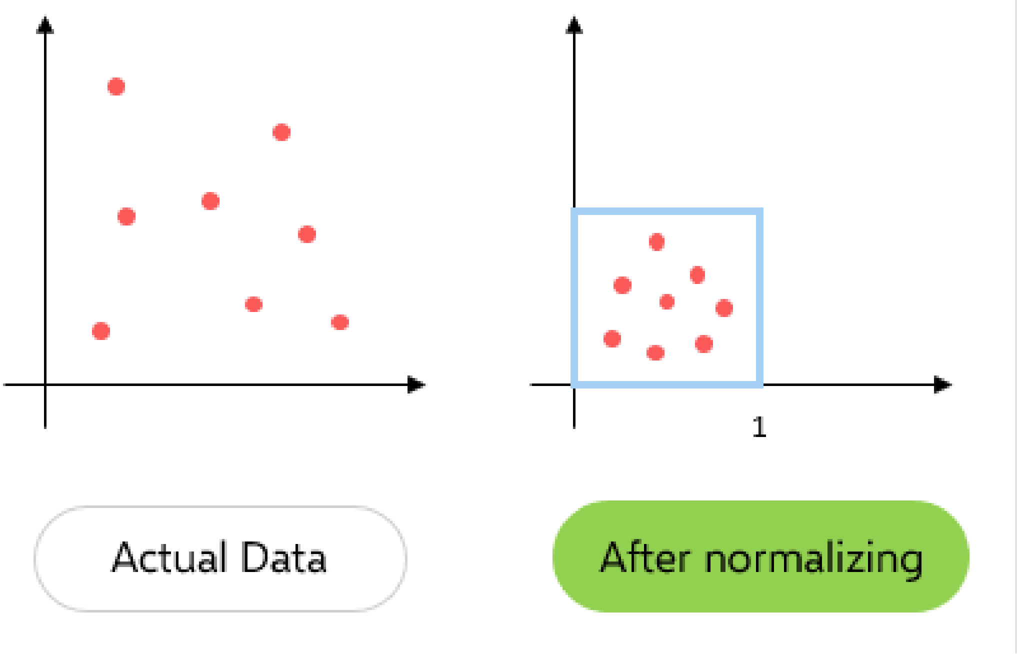

Figure 4 Visualisation of the effect of data normalisation (Richman, 2025)

Encoding Cyclical Features

Cyclical encoding was employed to represent repeating patterns in the categorical features in the dataset. Traditional methods like one-hot or ordinal encoding fail to reflect cyclic properties.

In a 24-hour clock, 23:00 and 00:00 are only one hour apart, yet ordinal encoding treats them as distant values. Cyclical encoding overcomes this limitation by transforming values into two dimensions using sine and cosine functions. This transformation maps cyclical values onto a circular space (see Figure 5), preserving their continuity and enabling models to recognise their repeating nature (Pelletier, 2025).

Using cyclical encoding improves machine learning models by preserving the relationships between cyclical values, helping them recognise time-based patterns more effectively. This technique is especially valuable for forecasting and anomaly detection since it preserves their cyclical significance for any analysis involving time or seasonal trends (Kud, 2023).

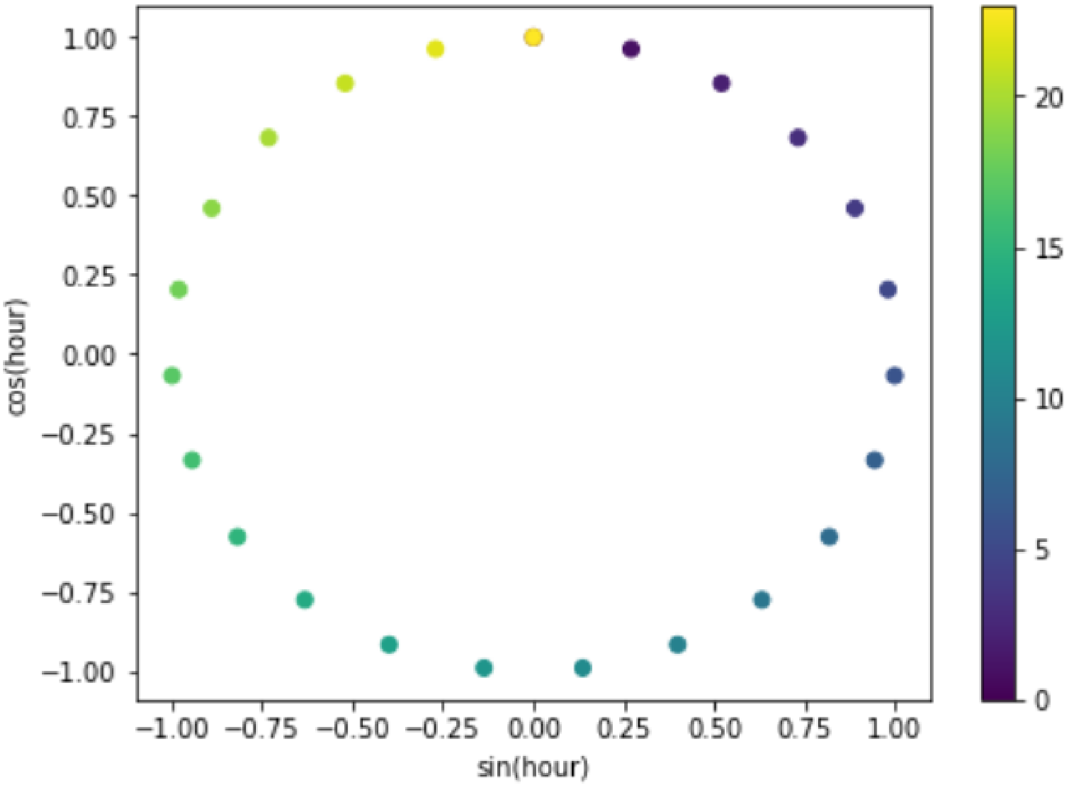

Figure 5 Visualisation of the effect of cyclical encoding (Cyclical Features, n.d.)

Descriptive Statistics

Table 2: Descriptive Statistics

Variable	Mean	Median	Std Dev	Min	Max
Energy production (kWh)	191.68	154.42	159.59	0.01	524.41
Meteorological Features					
Average environment temp (℃)	20.13	20.0	3.87	9.66	30.41
Average 1-second filtered wind speed(m/s)	6.81	6.39	3.57	0.01	17.02
Power Curtailment Features					
Average nacelle temp (℃)	29.06	28.96	3.84	18.7	39.21
Average rotor speed (rpm)	9.88	10.01	2.15	0.94	13.16
Energy production (kWh)	191.68	154.42	159.59	0.01	524.41

Average blade angle (°)	1.27	-0.01	3.32	-3.67	26.53
Yaw duration(s)	14.32	0.0	40.32	0.0	600.0
Turbine Age (months)	68.32	67.0	34.39	11.0	131.0
Interaction Features					
Wind Speed & Blade Angle	13.05	-0.02	36.66	-56.28	420.04
Rolling Features					
Wind Rolling Mean 24hr	6.81	6.47	2.4	0.83	15.4
Wind Rolling Std 24hr	2.49	2.44	0.95	0.0	5.94
Lagged Features					
Wind Speed 12hr Lag	6.81	6.39	3.57	0.01	17.02
Wind Speed 24hr Lag	6.81	6.39	3.57	0.01	17.02
Energy Production 12hr Lag	191.64	154.42	159.56	0.01	524.41
Energy Production 24hr Lag	191.6	154.42	159.53	0.01	524.41

The descriptive statistics of the wind turbine data offer valuable insights into the operational characteristics and environmental conditions influencing energy production.

Energy production exhibits considerable variability, with a mean of 191.68 kWh and a wide range extending from 0.01 kWh to 524.41 kWh. The high standard deviation of 159.59 kWh reflects substantial fluctuations in power output, likely driven by variations in wind speed and turbine efficiency. Such variability underscores the need for adaptive operational strategies to maintain more consistent energy generation. Lagged variables for both wind speed and energy production, measured at 12-hour and 24-hour intervals, display minimal deviation from current values. This stability indicates robust temporal autocorrelation, which is essential for short-term forecasting models, as historical data can reliably inform future predictions.

Wind speed, measured with a one-second filter, has an average of 6.81 m/s and a significant standard deviation of 3.57 m/s, ranging from near-calm conditions (0.01 m/s) to high wind speeds (17.02 m/s). The distribution reveals that most wind speeds fall between 4.09 m/s (25th percentile) and 9.2 m/s (75th percentile), suggesting a generally consistent wind resource interspersed with occasional high-speed gusts. Additionally, the 24-hour rolling mean of wind speed remains steady at 6.81 m/s, while the rolling standard deviation is 2.49 m/s, indicating a relatively stable wind pattern throughout the day. Nevertheless, the range of wind speeds (0.83 to 15.4 m/s) highlights the coexistence of both calm and gusty intervals. Furthermore, the interaction between wind speed and blade angle shows marked variability, with values ranging from -56.28 to 420.04. The positively skewed distribution implies that optimal combinations of high wind speed and certain blade angles may significantly enhance energy production.

Environmental conditions also play a crucial role in turbine performance. The average ambient temperature of 20.13°C, with a standard deviation of 3.87°C, reflects a moderate climatic environment, ranging from 9.66°C to 30.41°C. The nacelle temperature, with a mean of 29.06°C and a similar standard deviation of 3.84°C, is notably higher, indicating the impact of operational heat generation on the internal turbine environment. The difference between the nacelle and ambient temperatures underscores the importance of efficient temperature regulation to ensure turbine performance and longevity.

Rotor speed, averaging 9.88 rpm with a standard deviation of 2.15 rpm, demonstrates consistent rotational dynamics that correspond closely with wind speed variations. This alignment indicates that the turbine efficiently adapts to changing wind conditions to optimise power output. Additionally, the average blade angle of 1.27° shows substantial variability (standard deviation of 3.32°), ranging from -3.67° to 26.53°. This wide range reflects dynamic pitch adjustments, which are vital for maintaining aerodynamic efficiency, particularly during fluctuating wind conditions.

Yaw duration, with an average of 14.32 seconds and a high standard deviation of 40.32 seconds, exhibits substantial variation, with the maximum duration reaching 600 seconds. The yaw mechanism, which aligns the turbine with the prevailing wind direction, may experience prolonged adjustment periods due to unfavourable wind shifts or potential maintenance issues.

Model Selection

Regression-based machine learning models were preferred over time-series forecasting techniques due to their ability to capture complex, nonlinear relationships without requiring strict assumptions about stationarity. Unlike traditional time-series models, which often struggle to incorporate multiple exogenous variables effectively, machine learning regression-based approaches seamlessly integrate external factors like weather patterns, which enhance predictive accuracy. Scalability and computational efficiency further reinforce the advantage of machine learning models, as opposed to time-series methods such as ARIMA, which tend to become computationally expensive and require frequent retraining. This paper compares two machine learning models whilst utilising an Ordinary Least Squares (OLS) regression as a benchmark for comparison.

The OLS regression is a widely accepted and well-understood statistical method which provides a straightforward approach for initial analysis, offering clear insights into the relationship between independent and dependent variables. The coefficients obtained from an OLS regression provide a direct interpretation of each feature's influence on the outcome, contributing to an initial understanding of the dataset. Thus, the model serves as an appropriate benchmark due to its interpretability and lack of computational complexity.

The following machine learning models were selected for this research paper:

K-Nearest Neighbour

K-Nearest Neighbours (KNN) is a non-parametric, instance-based supervised learning algorithm applicable to both classification and regression tasks (Chakole, 2024). It operates on the premise that similar data points are typically located within proximity. In regressions, KNN predicts values by averaging the values of its nearest neighbours, with K representing the number of neighbours considered. The algorithm's predictive accuracy depends heavily on the selection of K and the proper normalisation of the model's features. Prior literature has assessed KNN's ability to predict wind power generation (Khujaev et al., 2024; Dhungana, 2025). Khujaev et al. (2024) found this machine learning algorithm to be a reliable technique for such predictions.

Decision Tree

Decision trees are a widely used machine learning model in studies focused on predicting renewable energy (Ledmaoui et al., 2023; Maghraoui et al., 2022). A Decision Tree is a machine learning algorithm used for both regression and classification tasks. A decision tree recursively splits the data based on significant input features until each subset is as pure as possible. Each split forms a decision node, and each leaf node represents a predicted outcome. The final prediction is made by tracing the path from the root to a leaf and averaging the target values of the instances that reach that leaf. The effectiveness of the model is dependent on data quality and hyperparameter selection (Mohapatra, 2023).

Random Forest Model

Random Forest regression is a robust ensemble learning method, well-suited for wind energy forecasting. Aggregating multiple decision trees enhances accuracy, stability, and resilience against overfitting - a common issue in individual trees sensitive to noise. This approach is particularly effective in managing fluctuations in wind speed and environmental variables (Malakouti, 2023; Aslan & Çelebi, 2024). A key advantage of Random Forest lies in its capacity to efficiently process high-dimensional datasets with minimal preprocessing.

XGBoost Model

Extreme Gradient Boosting (XGBoost) is a highly advanced machine learning algorithm known for its accuracy, efficiency, and robustness in wind energy forecasting. Its ability to handle noisy, high-dimensional data makes it well-suited for modelling the complex interactions evident in wind energy production. XGBoost's advanced boosting techniques iteratively refine weak learners, resulting in exceptional predictive performance. Additionally, the inclusion of L1 and L2 regularisation helps prevent overfitting while improving generalisation to unseen data (Wang et al., 2024). Previous studies have demonstrated that XGBoost surpasses Random Forest in predictive accuracy, making it a superior choice for wind energy forecasting (Imani et al., 2025).

Additionally, these models were selected over other machine learning models due to their significantly faster training time when handling this large dataset.

Train-test split

The dataset was split into input features and the target variable, energy production. To separate the data into training and testing subsets, the train_test_split function from the scikit-learn library was employed. The random_state parameter was utilised to maintain consistency in the data splitting process, ensuring reproducibility. This was essential for obtaining consistent results, which in turn supported the validity of the machine learning models' interpretations. Eighty percent of the data was designated for training, while the remaining twenty percent was allocated for model evaluation. The 80:20 split is a widely adopted practice in the machine learning domain, striking an optimal balance between training and testing data, thereby mitigating the risks of overfitting and underfitting (Ahmed, 2023). Testing on an independent dataset ensures that the models generalise effectively and serve as a benchmark for evaluating their reliability and appropriateness for this study.

Hyperparameter tuning

Randomised search was employed for hyperparameter tuning of the machine learning models. This method was selected due to its considerable reduction in computation time, its capacity to explore a wide range of hyperparameters, and its greater likelihood of identifying optimal configurations. It proves especially advantageous in situations with limited computational resources, as it efficiently samples hyperparameter combinations at random (Bergstra & Bengio, 2012; Hestisholihah, 2023). Model predictive accuracy is improved through the randomness inherent in hyperparameter tuning, which minimises the potential overfitting of the training data. The predictive performance of the machine learning models was evaluated based on the hyperparameter-tuned versions.

Model Evaluation & Interpretability

Model predictive performance

Cross-validation was used to assess the model's mean squared errors, as it provides a more reliable and thorough evaluation of performance on unseen data. The structured division of the dataset into several folds helps reduce variability in performance estimates and ensures that all data points are involved in the evaluation process (GeeksforGeeks, 2025a). Consequently, cross-validation stands as an indispensable tool in machine learning, ensuring rigorous and insightful model assessment and comparison.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used to evaluate the machine learning models and are calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MSE is a fundamental metric for assessing model performance, as it quantifies the average squared differences between predicted and actual values. The squaring process magnifies the impact of larger errors, making MSE particularly sensitive to outliers. As a result, models with lower MSE are deemed more accurate, as they exhibit smaller deviations from true values, while higher MSE values indicate greater prediction errors and reduced performance (GeeksforGeeks, 2020b).

RMSE, derived by taking the square root of MSE, serves as a similarly valuable evaluation metric. While RMSE provides a more interpretable scale, it retains MSE’s sensitivity to outliers by assigning greater weight to larger errors. Consequently, a lower RMSE signifies improved model accuracy, with predictions closely aligning with actual values, whereas a higher RMSE reflects greater discrepancies and diminished predictive performance (Chugh, 2020).

Furthermore, the models were evaluated using Mean Absolute Error (MAE) and R-squared (R^2) as key metrics.

MAE is a metric that quantifies the average absolute difference between predicted and actual values. It reflects the average magnitude of errors in predictions, offering an intuitive measure of model accuracy (Chugh, 2020). Lower MAE values indicate better predictive performance, and the metric is easily interpretable since it is expressed in the same units as the data.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

R^2 is a statistical measure that assesses how well the predicted values from a model align with the actual data. R^2 ranges from 0 to 1, where 1 indicates a perfect fit, and 0 suggests that the model explains none of the variance in the dependent variable. Essentially, R^2 indicates the proportion of variance in the dependent variable that is accounted for by the model, providing insight into its explanatory power (Chugh, 2020).

Model interpretability: SHAP

The model-agnostic approach, SHAP, was utilised instead of post-hoc techniques because of its flexibility and the interpretability it offers as an explainable artificial intelligence method. These methods worked well across different machine learning models while offering clear insights into prediction reasoning. A key advantage was their ability to generate counterfactual explanations. This minimises the black-box nature associated with machine learning models since the model-agnostic approach offers insight into the decision-making strategy employed. Additionally, their robustness across various data types makes them ideal for renewable energy research.

SHAP (SHapley Additive Explanations) is a model-agnostic explainability technique derived from cooperative game theory. It assigns an importance value to each feature by assessing its contribution to a specific prediction, ensuring a fair and consistent distribution of the model's output across all input features. Unlike other attribution methods, SHAP guarantees consistency and local accuracy, meaning that if a feature has a greater impact on a prediction, its assigned value will never be lower than that of a less influential feature (Molnar, 2025). By evaluating all possible feature combinations, SHAP provides a comprehensive and unbiased understanding of how each input affects the model's outcome, making it a powerful tool for both global and local interpretability (Huang et al., 2020, Awan, 2023). This capability is particularly useful in wind energy planning, where understanding the overall behaviour of predictive models and the role of individual features, such as wind speed and ambient temperature, is crucial for optimising turbine placement and forecasting energy yield.

In this study, SHAP was applied to the two best-performing models to enhance global interpretability and provide insights into individual feature contributions. By quantifying the impact of each feature on predictions, SHAP promotes transparency and accountability in the machine learning process (Molnar, 2025). Its application to the full dataset for both models identified the most influential features driving the models' decision-making, offering valuable guidance for policymakers and energy planners. This information allows for the prioritisation of key features in regulatory frameworks, reducing uncertainty in energy production estimates and ultimately supporting data-driven decision-making in wind energy development.

SHAP is highly informative but can be computationally demanding, especially for large datasets or complex models, which may limit its scalability in real-time applications. While it provides valuable global insights, it may sometimes overlook critical feature interactions that influence decision-making. Despite the limitations, SHAP offers actionable intelligence for decision-makers in wind energy planning, including policymakers, energy planners, and investors.

4. Empirical Analysis

Comparison of Machine Learning Models

The performance of the machine learning models is compared in Figures 6 and 7 below. The graphs provide insights into the predictive ability of the models based on the RMSE and MSE evaluation metrics.

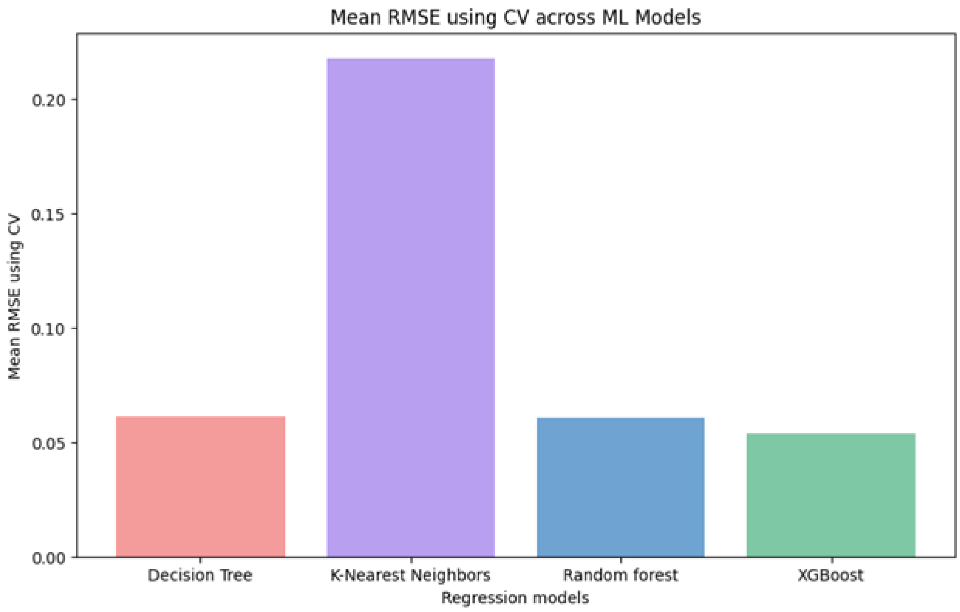

Figure 6 Comparison of mean RMSE using CV Across Machine Learning models

As shown in Figure 6, XGBoost outperforms the other machine learning models, achieving the lowest mean RMSE through cross-validation. KNN performs the worst, with a high mean RMSE above 0.2, while Random Forest and Decision Tree exhibit similar mean RMSE values, with Random Forest slightly outperforming the Decision Tree model.

Figure 7 provides a further comparison of mean MSE trends from cross-validation over 10 iterations. Consistent with the results in Figure 6, XGBoost consistently achieves a lower mean MSE across all iterations, demonstrating its robustness compared to the other machine learning models. In contrast, Random Forest and Decision Tree show similar mean MSE values throughout the iterations, while KNN displays a significantly higher mean MSE trend.

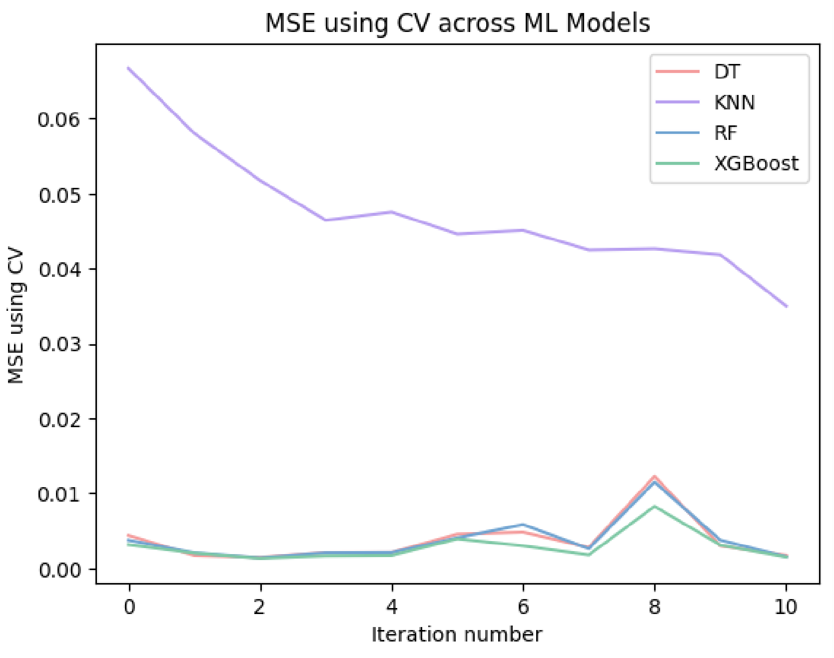

Figure 7 Comparison of MSE using CV across Machine Learning models

Table 3: Comparison of Model Performance

	Baseline Model	Machine Learning Models			Improvement from Baseline Model (%)		
Performance Metrics	OLS	XGBoost	Random Forest	Decision Tree	XGBoost	Random Forest	Decision Tree
Mean Absolute Error (MAE)	0.0849	0.0121	0.0187	0.0225	85.75%	77.97%	73.50%
Mean Squared Error (MSE)	0.1430	0.0010	0.0018	0.0023	99.30%	98.74%	98.39%
Root Mean Squared Error (RMSE)	0.1198	0.0319	0.0430	0.0484	73.37%	64.11%	59.60%
Coefficient of Determination (R^2)	0.8450	0.9891	0.9802	0.9748	17.05%	16.00%	13.32%

The MAE, MSE, and RMSE values for the XGBoost, Random Forest and Decision Tree machine learning models are significantly lower than those obtained from the OLS regression, demonstrating their superior ability to capture complex relationships within the data.

Among the machine learning models assessed, XGBoost emerges as the most effective forecasting tool, consistently outperforming the Random Forest and Decision Tree models across all key performance metrics. While the Decision Tree model remains highly effective, it falls slightly behind Random Forest in all evaluation metrics, underscoring Random Forest's higher predictive performance in comparison. The high R^2 values across the models raise concerns about overfitting. However, since hyperparameter tuning has been performed, the values likely represent optimised model performance.

The percentage improvements further highlight this advantage, with XGBoost achieving a significant 99.30% reduction in MSE and a 73.37% reduction in RMSE when compared to the baseline model (OLS). These findings underscore the critical role of machine learning regression models in accurately forecasting energy production, particularly in capturing non-linear and dynamic data patterns.

OLS Regression

Table 4: OLS Regression Results

Variable	Coefficient	Std Error	t	P > |t|	(0.025	0.975)
Intercept	-0.690	0.001	-560.33	0.000	-0.692	-0.687
Average Rotor Speed	0.995	0.001	674.458	0.000	0.992	0.998
Average Filtered Wind Speed	0.390	0.002	258.270	0.000	0.387	0.393
Wind Speed x Blade Angle	0.374	0.005	69.752	0.000	0.364	0.385
Average Blade Angle	0.332	0.003	99.817	0.000	0.326	0.339
Average Nacelle Temp	0.067	0.001	49.728	0.000	0.065	0.070
Yaw Duration	0.061	0.003	23.217	0.000	0.056	0.066
Wind Rolling Std 24hr	0.058	0.001	48.731	0.000	0.055	0.060
Energy Production 12hr Lag	0.052	0.001	51.587	0.000	0.050	0.054
Wind Speed 12hr Lag	-0.043	0.002	-25.386	0.000	-0.046	-0.040
Wind Speed 24hr Lag	-0.043	0.001	-28.769	0.000	-0.045	-0.040
Wind Rolling Mean 24hr	0.038	0.002	16.744	0.000	0.034	0.042
Energy Production 24hr Lag	0.020	0.001	20.318	0.000	0.018	0.022
Average Environment Temp	-0.010	0.001	-7.334	0.000	-0.012	-0.007
Month Sin	-0.008	0.001	-13.391	0.000	-0.009	-0.007
Hour Cos	0.007	0.000	25.981	0.000	0.006	0.007
Hour Sin	0.003	0.000	12.798	0.000	0.003	0.004
Month Cos	-0.003	0.001	-4.833	0.000	-0.004	-0.002
Season Sin	-0.003	0.001	-4.177	0.000	-0.004	-0.001
Day of Month Sin	0.002	0.000	6.358	0.000	0.001	0.002
Day of Month Cos	-0.001	0.000	-2.851	0.004	-0.001	0.000
Season Cos	-0.001	0.001	-1.236	0.217	-0.002	0.000
Turbine Age	0.001	0.001	0.753	0.451	-0.001	0.002
Day of Week Cos	0.000	0.000	1.805	0.071	0.000	0.001
Day of Week Sin	0.000	0.000	0.839	0.402	0.000	0.001

The estimated regression coefficients offer valuable insights into the primary factors influencing wind energy production (kWh). Among the most significant predictors, average rotor speed ($\beta = 0.995$, $p < 0.001$) emerges as the primary driver of energy output in this model, exhibiting an almost one-to-one, statistically significant relationship. This indicates that a 1-unit increase in rotor speed (measured in revolutions per minute) results in an approximate 0.995 kWh increase in energy production, highlighting the critical role of rotor dynamics in optimising efficiency. The confidence interval for rotor speed ranges from 0.992 to 0.997, indicating that the estimate is highly precise.

Similarly, average filtered wind speed ($\beta = 0.390$, $p < 0.001$) serves as a fundamental determinant of energy production, reinforcing the well-established principle that higher wind speeds lead to greater power generation. The confidence interval for wind speed ranges from 0.387 to 0.393, further demonstrating strong precision and reliability in this estimate. Unlike rotor speed, which responds mechanically to wind conditions, filtered wind speed reflects the aerodynamic forces acting on the turbine. The coefficient suggests that for every 1 m/s increase in wind speed, energy output rises by 0.39 kWh, underscoring the importance of strategic site selection for wind farms.

Blade-related factors also significantly influence wind energy production. The average blade angle ($\beta = 0.332$, $p < 0.001$) demonstrates that optimising the orientation of turbine blades enhances aerodynamic efficiency, thereby increasing energy output. The confidence interval for blade angle ranges from 0.326 to 0.339, further underscoring the precision and reliability of this estimate. This aligns with the fundamental principle of maximising lift while minimising drag on the turbine blades, ensuring that turbines extract the maximum possible kinetic energy from the wind.

The expected relationship between yaw duration and power generation is inverse: the longer the yaw duration, the less efficiently the turbine generates power. Ideally, a shorter yaw duration, indicating quicker alignment with the wind, maximises the turbine's power generation potential. However, the model outputs an opposite relationship for yaw duration ($\beta = 0.061$, $p < 0.001$), suggesting a positive correlation. This could be due to several factors, such as omitted variables, measurement errors, or reverse causality, where other unaccounted-for factors may be influencing both yaw duration and power generation. Additionally, non-linear dynamics or model specification issues could also lead to this unexpected result.

The average nacelle temperature ($\beta = 0.067$, $p < 0.001$) shows a positive correlation with energy production. This relationship may be attributed to higher nacelle temperatures, indicating the consistent operation of the components within the nacelle, which in turn raises the temperature. However, it is important to note that exceeding a certain threshold in nacelle temperature could prove detrimental to the internal components due to thermal throttling.

Conversely, the average environmental temperature ($\beta = -0.010$, $p < 0.001$) has a slight negative effect. This is likely due to the inverse relationship between air temperature and density, where higher temperatures result in less dense air, which reduces the available kinetic energy in the wind.

Temporal dependencies in wind energy production reveal notable patterns. Energy production lagged by 12 hours ($\beta = 0.052$, $p < 0.001$) suggests that previous energy output has a small yet persistent influence on current production, indicating that stable wind conditions contribute to sustained power generation. However, wind speed lagged by 12 hours ($\beta = -0.043$, $p < 0.001$) negatively impacts energy production, likely reflecting the effects of turbine curtailment or mechanical strain following periods of strong winds. A similar trend is observed for wind speed lagged by 24 hours ($\beta = -0.043$, $p < 0.001$), reinforcing the notion that prolonged exposure to extreme wind speeds may disrupt continuous energy generation, either through temporary shutdowns to prevent overloading or increased fatigue on turbine components.

Cyclically encoded time variables provide important insights into the periodic nature of wind energy production in South Africa, capturing both daily and seasonal fluctuations in wind patterns specific to the region. The hour cosine ($\beta = 0.007$, $p < 0.001$) and hour sine ($\beta = 0.003$, $p < 0.001$) indicate that energy production follows a clear diurnal cycle, with fluctuations in output over 24 hours. The positive coefficient for the hour cosine term suggests that wind energy production is generally higher at certain times of the day, likely corresponding to peak wind periods, such as midday or late evening, when coastal winds are typically stronger. The smaller but significant hour sine component suggests there may be an asymmetric daily pattern, with certain times of day experiencing more favourable wind conditions, as is common in South Africa's coastal areas due to land-sea temperature differences.

Seasonal variability is reflected in the month sine ($\beta = -0.008$, $p < 0.001$) and month cosine ($\beta = -0.003$, $p < 0.001$) coefficients, indicating moderate fluctuations in wind energy production throughout the year. The negative coefficient for the month sine variable suggests that energy production tends to be lower during the months of December to February, when wind speeds may be less consistent. In contrast, production tends to be higher during the months of winter and early spring in South Africa, from June to August, which corresponds to stronger and more frequent winds. The month cosine term, while smaller in magnitude, further reflects seasonal variability, capturing gradual increases or decreases in wind patterns throughout the year. However, the season cosine ($\beta = -0.001$, $p = 0.217$) is statistically insignificant, suggesting that broader seasonal effects do not uniformly impact wind energy production.

Several variables exhibit statistical insignificance, indicating limited influence on wind energy production. Turbine age ($\beta = 0.001$, $p = 0.451$) suggests that, when adequately maintained, turbines do not experience a significant decline in efficiency over time. Similarly, day-of-week cyclic components ($p > 0.05$) are not significant. These findings underscore the reliability and consistency of wind power generation, which remains largely unaffected by short-term operational cycles or arbitrary time-based trends.

The low standard errors observed across all variables in the model suggest that the coefficient estimates are both precise and stable, instilling confidence that the true values closely align with the estimates. The OLS results highlight the dominant role of rotor speed, wind speed, blade angle, and yaw control in determining wind energy production. While temporal and cyclical fluctuations influence the output, aerodynamic and meteorological factors primarily drive energy production.

The model demonstrates a robust fit, with an R-squared value of 0.845, indicating that the model accounts for 84.5% of the variance in energy production. The equivalent R-squared and adjusted R-squared confirm the relevance of all input features. The high F-statistic of $1.114e^5$ with a near-zero p-value validates the statistical significance of the model and the importance of the predictors.

Despite the strong fit, residual analysis reveals potential issues. The Durbin-Watson statistic of 0.675 indicates positive autocorrelation, suggesting residuals are not entirely independent—a common issue in time series data. Additionally, the Omnibus and Jarque-Bera tests point to non-normality in the residuals, which may compromise the reliability of hypothesis testing and confidence intervals.

The condition number of 67.3 confirms the absence of significant multicollinearity, indicating that the predictors are not highly correlated. Nonetheless, the presence of autocorrelation and non-normal residuals suggests that more advanced modelling techniques should be explored to enhance the predictive accuracy and reliability of the model.

Random Forest and XGBoost

To achieve an average explanation across the dataset, SHAP analysis was applied to both XGboost and Random Forest machine learning models. For each model, a SHAP summary plot illustrating the top five contributing features was created. Figures 8 and 9 present the global interpretation SHAP summary plots for these models, where the SHAP value is measured in log odds.

The top five features that appeared in both models include: Average rotor speed (rpm), Average blade angle (°), Average 1-second filtered wind speed (m/s), Interaction of Wind Speed & Blade Angle and Turbine Age. The recurring presence of these features in both models emphasised their consistent influence on wind energy production predictions. This consistency indicated that these features were not model-specific but were essential in predicting the total energy production of wind turbines (kWh).

The contribution of these features extended across the entire dataset, ensuring global interpretability. Their strong predictive power at a global level indicated that variations in these features significantly impacted the model's output. As a result, these factors were essential in understanding and enhancing prediction accuracy.

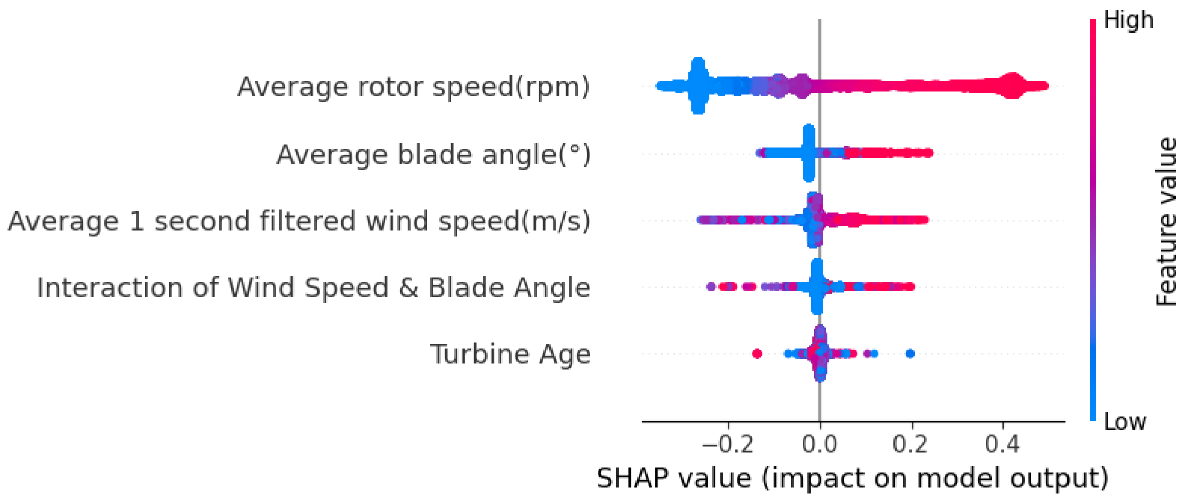

Figure 8 SHAP Summary Plot (Feature Impact on Random Forest Predictions)

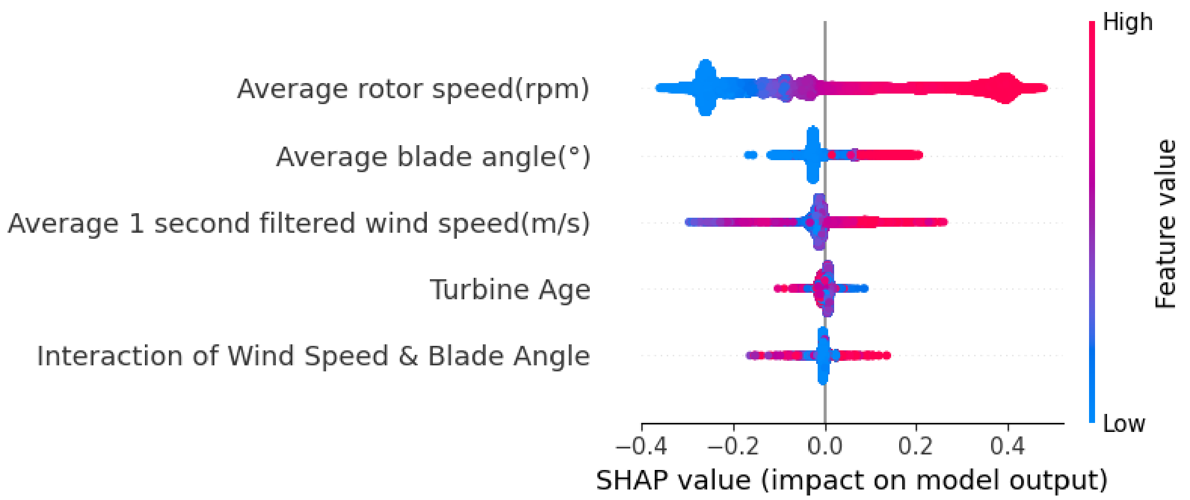

Figure 9 SHAP Summary Plot (Feature Impact on XGBoost Predictions)

Both models' SHAP summary plots reveal that average rotor speed (rpm) significantly influences energy production. Specifically, higher rotor speeds increase the log odds by 0.45, which raises the probability of higher energy production from 50% to approximately 61%. In contrast, lower rotor speeds decrease the log odds by -0.35, lowering the probability to around 41.3%. Notably, this effect is asymmetric, with higher rotor speeds exerting a stronger influence on energy production than lower speeds.

Similarly, both the Random Forest and XGBoost SHAP summary plots indicate that larger blade angles increase the log odds by 0.3, which boosts the probability of higher energy production by approximately 7.5%. Conversely, smaller blade angles reduce the log odds by -0.18, decreasing the probability by 4.5%. This highlights the asymmetric role of blade positioning, with larger angles having a more pronounced effect on energy production.

It is evident from both plots that higher wind speeds, in general, increase the log odds by 0.28, raising the probability of higher energy production to 57%. However, there are instances of average to high wind speeds where the probability of increased energy production diminishes, likely due to the negative impact of high wind variability.

Figures 8 and 9 illustrate a reversal in the feature ranks between the interaction of wind speed and blade angle, and the turbine age feature. Despite the shift in feature importance rank, both SHAP summary plots demonstrate similar impacts on the model. For the interaction term, a weak interaction between wind speed and blade angle shows a minimal to negligible effect on energy production, while a stronger interaction generally leads to an increase in energy production, though average interaction values may result in a decrease in the probability of increasing wind energy production.

In the Random Forest model, the SHAP values for turbine age are centred around 0, indicating a minimal to negligible effect on the probability of affecting wind energy production. This could suggest that regular maintenance, such as the replacement or repair of worn-out components, mitigates any adverse effects of turbine age. While a similar trend is observed in the XGBoost model, there is a clearer expected relationship, that a higher turbine age decreases the probability of increasing energy production by approximately 2.5%, and vice versa.

5. Discussion

The results of the analysis provide compelling evidence that machine learning models, particularly XGBoost, yield significantly more accurate forecasts under local conditions when compared to other machine learning models and the traditional statistical method OLS. The performance metrics from Table 3 indicate that the machine learning models drastically reduce forecasting errors while improving predictive accuracy.

The substantial performance gap between OLS and XGBoost, particularly the near-total elimination of squared errors, highlights the capability of gradient-boosting algorithms to refine predictions through iterative learning. This suggests that for real-world applications requiring precision, such as wind energy forecasting, machine learning methods, especially the ensemble model XGBoost, are far superior to conventional regression techniques, which is consistent with prior research (Imani et al., 2025).

The demonstrated superiority of machine learning techniques suggests that reliance on traditional statistical models may lead to suboptimal predictions. By leveraging advanced algorithms like XGBoost, stakeholders can significantly enhance forecasting accuracy, leading to better resource allocation, improved grid stability, and more efficient energy production planning.

The SHAP summary plots (see Figures 8 and 9) reveal key determinants of wind energy production. Notably, critical operational variables include average rotor speed, average blade angle, and turbine age, while the essential meteorological factor is the average 1-second filtered wind speed. Additionally, the interaction between an operational and meteorological factor, namely wind speed and blade angle, exerts a significant influence on energy output.

Average rotor speed is a critical operational variable which plays a pivotal role in wind energy production. The rotor is essential as it converts wind into electricity through aerodynamic forces. Higher rotor speeds increase the probability of greater energy output, while lower speeds diminish it. This is expected, as rotor speed directly affects the kinetic energy conversion efficiency of the turbine. An asymmetric effect is observed from the machine learning models, where increases in rotor speed have a stronger impact than decreases in energy production. To optimise wind energy production, dynamic control systems should be employed, along with maintenance and real-time monitoring to maintain rotor speed within an optimal range, ensuring maximum efficiency while preventing mechanical wear.

Blade angle also exerts a substantial influence on wind energy production. The model results show that larger angles increase energy output by enhancing the aerodynamic efficiency of the blades, whereas smaller angles reduce production. This aligns with the fundamental principle of maximising lift while minimising drag on the turbine blades, ensuring that turbines extract the maximum possible kinetic energy from the wind. The asymmetry relationship suggests that optimising blade positioning is essential for maximising efficiency. Advanced adaptive pitch control mechanisms, integrated with real-time wind monitoring systems, can be implemented to continuously adjust blade angles for optimal energy capture. Regular inspections and automated sensors should also be utilised to ensure that the blade angle remains within the optimal range, minimising performance losses due to mechanical degradation or misalignment.

The XGBoost model suggests that as the wind turbine ages, there is a slight decline in energy production. This declining relationship is expected and is likely due to wear and tear. However, the minimal nature of this decline indicates that periodic maintenance of the wind turbine has been performed. Implementing predictive maintenance strategies, supported by sensors, can be employed to detect early signs of performance degradation and schedule necessary repairs to further minimise energy loss.

The interaction between wind speed and blade angle also plays a crucial role in determining overall energy output, as per the model’s results. While results show weak interactions have minimal effects, stronger interactions generally lead to increased production. This relationship is expected, as lift depends on the blade angle relative to the incoming wind flow and is directly proportional to wind speed.

Wind speed is a fundamental meteorological factor affecting wind energy production, with the models showing that high speeds generally increase output. However, variability in wind speed can lead to a reduction in efficiency, likely due to turbulence and rapid fluctuations disrupting the turbine's stability. Therefore, careful site selection is crucial, favouring locations with consistently strong and stable winds, such as coastal areas or elevated terrains. Additionally, integrating energy storage solutions can mitigate the impact of wind variability, ensuring a steady supply of electricity even when wind conditions fluctuate.

For policy interventions, the aforementioned key insights are essential, as they highlight the significant impact of maintenance and site selection on the overall performance of wind turbines. Any policy measures aimed at improving wind turbine maintenance will directly contribute to increased wind energy production in wind farms.

To improve turbine maintenance and minimise operational downtime, policies should focus on incentivising local manufacturing of turbine components. According to a WWF SA article, only towers, transformers, and certain cabling for wind turbines are currently produced locally (*Opportunities in South Africa’s Wind Energy Value Chain*, 2022). Expanding domestic production would not only create jobs but also shorten maintenance lead times and reduce reliance on imported components. Additionally, lowering import duties and taxes on essential turbine parts could further reduce maintenance costs, enhancing the economic viability of wind energy production.

Establishing standardised maintenance protocols is a crucial step in ensuring long-term efficiency and minimising unexpected energy losses. Policies should mandate preventative maintenance as an industry standard to sustain optimal turbine functionality. This includes the implementation of automated performance monitoring systems, such as vibration sensors on rotor components, to detect early signs of mechanical degradation.

Beyond maintenance, policies should promote investment in battery storage technologies to address the variability in wind energy production caused by fluctuations in wind speed. To ensure a stable power supply, policies that offer incentives for hybrid renewable energy projects, such as integrating wind energy with battery storage, can enhance grid stability and reliability.

Furthermore, developing a skilled workforce is vital to supporting South Africa's growing wind energy sector. Policy intervention focused on investment in specialised training programs for wind turbine technicians, along with government funding for technical schools and engineering programs focused on wind energy, will help build a sustainable pipeline of skilled professionals.

The inclusion of the explainable machine learning technique, SHAP, clarifies how machine learning models generate predictions by quantifying each feature's contribution. This allows policymakers and grid operators alike to identify key variables enhancing model transparency and ensuring more reliable decision-making. Its intuitive visualisations make machine learning models more accessible, bridging the gap between AI forecasting and practical energy applications. This transparency fosters greater adoption of advanced forecasting tools, ultimately improving the reliability and integration of machine learning models in the energy sector.

6. Conclusion

This study addressed a central research problem: evaluating the effectiveness of machine learning models in forecasting wind energy production in South Africa, a country marked by highly variable geographic and climatic conditions. With increasing investment in renewable energy, particularly wind power, understanding which predictive tools offer the greatest accuracy and operational insight is critical. The research also sought to determine which operational and meteorological variables most influence energy production and how the interpretability of models can be improved to enhance trust and adoption among stakeholders in the energy sector.

The analysis revealed that advanced machine learning models, particularly XGBoost, outperform traditional linear models and even other ensemble methods in predicting wind energy production. Random Forest emerged as a close second in terms of accuracy. These models consistently demonstrated superior performance based on standard evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and the R-squared statistic. Through explainable AI techniques such as SHAP values, the study identified five critical variables that consistently influenced energy output across models: average rotor speed, blade angle, filtered wind speed, turbine age, and the interaction between wind speed and blade angle. These findings not only align with theoretical expectations but also validate the relevance of these features in real-world turbine performance, underscoring their potential for guiding operational decisions.

The study employed a rigorous methodological approach using a dataset of 10-minute interval wind turbine data spanning 2015 to 2024. Extensive preprocessing was conducted to address missing data, outliers, and normalisation issues, ensuring a high-quality input for modelling. Time-series characteristics were handled through cyclical encoding and feature engineering, including lagged variables, rolling statistics, and interaction terms. The models were trained and validated using an 80/20 data split, and hyperparameters were tuned via RandomizedSearchCV. Cross-validation ensured the robustness and generalisability of the models, and a baseline comparison with OLS regression provided a benchmark for measuring the gains from more complex algorithms.

The findings of this study have significant practical and policy implications. The identification of key performance variables can inform maintenance scheduling, optimal turbine placement, and site selection strategies. For policymakers and energy planners, the results suggest that integrating explainable machine learning models into forecasting systems could enhance reliability and planning capacity in renewable energy integration. Furthermore, the consistent superiority of ensemble methods like XGBoost supports the case for their adoption in operational decision-making and energy management systems. At the policy level, these findings suggest the need to invest in infrastructure such as battery storage to accommodate the variability in wind power generation and to expand training programs for wind energy technicians to support the sector's growth.

Despite the promising results, the study is not without limitations. The analysis was based on data from a single wind turbine selected for its data availability rather than representativeness. This limits the generalisability of the findings to other wind farms or turbine types operating under different environmental or mechanical conditions. Additionally, the model did not include certain potentially important environmental variables, such as air pressure and humidity, due to data limitations. These omissions may have affected the accuracy and explanatory power of the models.

Future research should aim to extend this work by incorporating data from multiple turbines located in diverse regions across South Africa. This would allow for a broader assessment of model robustness and enable comparative studies across different climatic and topographic conditions. Furthermore, integrating additional environmental variables and experimenting with hybrid modelling approaches—combining physical and statistical models—could further enhance predictive accuracy and operational relevance. As the energy sector transitions toward greater reliance on renewables, such refined modelling efforts will be essential to ensuring reliability, efficiency, and equitable access.

7. Reference List

Ahmed. (2023, September 9). The motivation for Train-Test split - Ahmed - Medium. *Medium*. http://medium.com/@nahmed3536/the-motivation-for-train-test-split-2b1837f596c3

Aslan, A., & Çelebi, S. (2024). Improving Forecasting Performance with Random Forest Regression Model in Wind Power Generation. https://www.researchgate.net/publication/385168698_Improving_Forecasting_Performance_with_Random_Forest_Regression_Model_in_Wind_Power_Generation

Awan, A (2023, June 28). *An Introduction to SHAP Values and Machine Learning Interpretability*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(1), 281–305. https://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf

Bousla, M., Belfkir, M., Haddi, A., Mourabit, Y. E., & Bossoufi, B. (2024). Comparison of artificial intelligence approaches for estimating wind energy production: A Real-World Case Study. *Results in Engineering*, 103626. https://doi.org/10.1016/j.rineng.2024.103626

Burton, T., Jenkins, N., Sharpe, D., & Bossanyi, E. (2011). *Wind Energy Handbook*. https://doi.org/10.1002/9781119992714

Chakole, A. (2024, November 15). K-Nearest Neighbors (KNN) - Anup Chakole - Medium. *Medium*. https://medium.com/@anupchakole/k-nearest-neighbors-knn-d8013486db6f

Chen, Z., Shi, T., Song, P., Li, C., Cao, Y., & Yan, Y. (2023). Improved pitch control strategy for the robust operation of wind energy conversion system in the high wind speed condition. *International Journal of Electrical Power & Energy Systems*, *153*, 109381. https://doi.org/10.1016/j.ijepes.2023.109381

Chugh, A. (2020, December 8). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium*. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Cowling, N. (2024, April 2). *Share of electricity consumption in South Africa 2019, by sector*. Statista. https://www.statista.com/statistics/1221769/electricity-consumption-in-south-africa-by-sector/

CRSES. (2025, March 20). *SA Electricity Made Visual*. CRSES | the Centre for Renewable and Sustainable Energy Studies. https://www.crses.sun.ac.za/sa-energy-stats/

Cyclical Features. (n.d.). Feature Engine. https://feature-engine.trainindata.com/en/1.7.x/user_guide/creation/CyclicalFeatures.html

Daniel, L., Sigauke, C., Chibaya, C., & Mbuvha, R. (2020). Short-Term wind speed forecasting using statistical and machine learning methods. *Algorithms*, *13*(6), 132. https://doi.org/10.3390/a13060132

Dhungana, H. (2025). A machine learning approach for wind turbine power forecasting for maintenance planning. *Energy Informatics*, *8*(1). https://doi.org/10.1186/s42162-024-00459-4

Eberhard, A., & Naude, R. (2016). The South African Renewable Energy Independent Power Producer Procurement Programme: A review and Lessons learned. *Journal of Energy in Southern Africa*, *27*(4), 1. https://doi.org/10.17159/2413-3051/2016/v27i4a1483

Elsaraiti, M., Merabet, A., & Al-Durra, A. (2019). Time series analysis and forecasting of wind speed data. *IEEE Industry Applications Society Annual Meeting*, 1–5. https://doi.org/10.1109/ias.2019.8912392

GeeksforGeeks. (2025a, February 28). *Cross validation in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/cross-validation-machine-learning/

GeeksforGeeks. (2024b, July 3). *Evaluation metrics in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/metrics-for-machine-learning-model/

Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. *Energy Economics*, *81*, 709–727. https://doi.org/10.1016/j.eneco.2019.05.006

Hestisholihah. (2023, June 19). Hyperparameter Tuning Showdown: Grid Search vs. Random Search — Which is the Ultimate Winner? *Medium*. https://medium.com/@hestisholihah01/hyperparameter-tuning-showdown-grid-search-vs-random-search-which-is-the-ultimate-winner-5927b322e54d

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., & Yi, X. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, *37*, 100270. https://doi.org/10.1016/j.cosrev.2020.100270

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*, *13*(3), 88. https://doi.org/10.3390/technologies13030088

Jha, A. (2023, August 18). *ML Model Interpretation Tools: What, why, and how to interpret*. neptune.ai. https://neptune.ai/blog/ml-model-interpretation-tools

Khan, P. W., Byun, Y., Lee, S., & Park, N. (2020). Machine learning based hybrid system for imputation and efficient energy demand forecasting. *Energies*, *13*(11), 2681. https://doi.org/10.3390/en13112681

Khujaev, A., Yildirim, O., & Benzaibak, F. (2024). Estimating the wind power by using K-Nearest Neighbors (KNN) and Python. *Engineering and Technology Journal*, *09*(09). https://doi.org/10.47191/etj/v9i09.13

Krishnamurthy, S., Adewuyi, O. B., Luwaca, E., Ratshitanga, M., & Moodley, P. (2024). Artificial intelligence-based forecasting models for integrated energy system management planning: An exploration of the prospects for South Africa. *Energy Conversion and Management X*, *24*, 100772. https://doi.org/10.1016/j.ecmx.2024.100772

Kud, A. (2023, December 4). Why we need encoding cyclical features. *Medium*. https://medium.com/@axelazara6/why-we-need-encoding-cyclical-features-79ecc3531232

Lange, M., & Focken, U. (2006). Physical wind power Prediction Systems. In *Springer eBooks* (pp. 39–53). https://doi.org/10.1007/3-540-31106-8_4

Ledmaoui, Y., Maghraoui, A. E., Aroussi, M. E., Saadane, R., Chebak, A., & Chehri, A. (2023). Forecasting solar energy production: A comparative study of machine learning algorithms. *Energy Reports*, *10*, 1004–1012. https://doi.org/10.1016/j.egyr.2023.07.042

Li, X., Li, K., Shen, S., & Tian, Y. (2023). Exploring Time Series Models for Wind Speed Forecasting: A Comparative analysis. *Energies*, *16*(23), 7785. https://doi.org/10.3390/en16237785

Liu, X., Lin, Z., & Feng, Z. (2021). Short-term offshore wind speed forecast by seasonal ARIMA - A comparison against GRU and LSTM. *Energy*, *227*, 120492. https://doi.org/10.1016/j.energy.2021.120492

Maghraoui, A. E., Hammouch, F., Ledmaoui, Y., & Chebak, A. (2022). Smart Energy Management System: A Comparative Study of Energy Consumption Prediction Algorithms for a Hotel Building. *2022 4th Global Power, Energy and Communication Conference (GPECOM)*. https://doi.org/10.1109/gpecom55404.2022.9815807

Makubyane, K., & Maposa, D. (2024). Forecasting Short- and Long-Term wind speed in Limpopo Province using machine learning and extreme value theory. *Forecasting*, *6*(4), 885–907. https://doi.org/10.3390/forecast6040044

Malakouti, S. M. (2023). Estimating the output power and wind speed with ML methods: A case study in Texas. *Case Studies in Chemical and Environmental Engineering*, *7*, 100324. https://doi.org/10.1016/j.cscee.2023.100324

Mhlanga, D. (2023). Artificial intelligence and machine learning for energy consumption and production in emerging markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4319461

Mohapatra, R. (2023, February 21). A Step-by-Step Guide to building accurate Predictive Decision Tree Model. *Medium*. https://techwithram.medium.com/a-step-by-step-guide-to-building-accurate-predictive-decision-tree-model-598a5cfb460d

Molnar, C. (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3rd ed.). https://christophm.github.io/interpretable-ml-book

Olawade, D. B., Wada, O. Z., Ige, A. O., Egbewole, B. I., Olojo, A., & Oladapo, B. I. (2024). Artificial intelligence in environmental monitoring: advancements, challenges, and future directions. *Hygiene and Environmental Health Advances*, 100114. https://doi.org/10.1016/j.heha.2024.100114

Opportunities in South Africa's wind energy value chain. (2022). WWF South Africa. https://www.wwf.org.za/?41662/wind-energy-fact-sheet

Pelletier, H. (2025, January 21). *Cyclical Encoding: an alternative to One-Hot encoding for time series features*. Towards Data Science. https://towardsdatascience.com/cyclical-encoding-an-alternative-to-one-hot-encoding-for-time-series-features-4db46248ebba/

Ramarope, S. I., Fatoba, O. S., & Jen, T. (2023). Hydro-power generation forecast in South Africa based on Machine Learning (ML) models. *Scientific African*, *22*, e01981. https://doi.org/10.1016/j.sciaf.2023.e01981

Richman, J. (2025, January 30). Data Normalization explained: Types, examples, & methods. *Estuary*. https://estuary.dev/blog/data-normalization/

South African Wind Atlas. (2020). South Africa Wind Energy Association. https://sawea.org.za/south-african-wind-atlas

Szostek, K., Mazur, D., Drałus, G., & Kusznier, J. (2024). Analysis of the effectiveness of ARIMA, SARIMA, and SVR models in time series Forecasting: A case study of wind farm energy production. *Energies*, *17*(19), 4803. https://doi.org/10.3390/en17194803

Thanyani, M. (2020, August 12). *Forecasting hourly electricity demand in South Africa using machine learning models*. http://hdl.handle.net/11602/1595

Tyass, I., Bellat, A., Raihani, A., Mansouri, K., & Khalili, T. (2022). Wind Speed Prediction Based on Seasonal ARIMA model. *E3S Web of Conferences*, *336*, 00034. https://doi.org/10.1051/e3sconf/202233600034

Wang, R., Wu, J., Cheng, X., Liu, X., & Qiu, H. (2024). Adaptive expert fusion model for online wind power prediction. *Neural Networks*, *184*, 107022. https://doi.org/10.1016/j.neunet.2024.107022

Wright, M. A., & Grab, S. W. (2017). Wind speed characteristics and implications for wind power generation: Cape regions, South Africa. *South African Journal of Science*, *113*(7/8), 8. https://doi.org/10.17159/sajs.2017/20160270

Zhou, H., Liu, Q., Yan, K., & Du, Y. (2021a). Deep Learning Enhanced Solar Energy Forecasting with AI-Driven IoT. *Wireless Communications and Mobile Computing*, *2021*, 1–11. https://doi.org/10.1155/2021/9249387

Zhou, Z., Wang, S., Jiang, J., Li, H., Zhao, J., & Yu, D. (2025b). Optimizing wind turbine blade pitch control via input output differential model free adaptive control. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-88711-z